DOCUMENT RESUME

ED 389 310 IR 055 693

AUTHOR Micco, Mary; Popp, Rich

TITLE Developing an Information Infrastructure To Support Information Retrieval: Towards a Theory of Clustering Based in Classification.

SPONS AGENCY Council on Library Resources, Inc., Washington, D.C.; Department of Education, Washington, DC.

PUB DATE 94

NOTE 22p.; For a related journal article, see EJ 478 034.

PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Automatic Indexing; *Classification; College Libraries; Databases; Expert Systems; Higher Education; Hypermedia; *Information Retrieval; *Online Catalogs; *Search Strategies; *Subject Index Terms; Users (Information)

IDENTIFIERS *Cluster Based Retrieval; *Dewey Decimal Classification; Indiana University of Pennsylvania; Information Infrastructure; Library of Congress Subject Headings; MARC; Natural Language; Prototypes

ABSTRACT

Techniques for building a world-wide information infrastructure by reverse engineering existing databases to link them in a hierarchical system of subject clusters to create an integrated database are explored. The controlled vocabulary of the Library of Congress Subject Headings is used to ensure consistency and group similar items. Each database becomes a system object, and each package within the database is assigned a subject cluster based on its content. An expert system matches the user profile to the information package best suited to need and locates the appropriate database. This is supplemented by a machine-generated natural language mapping scheme to lead the user into the clusters of interest. For the prototype, an object-oriented hypermedia user interface was developed, using MARC records. Packages are grouped into subject clusters consisting of the classification number and the first subject heading/keyword assigned. Use of a hierarchical classification number (Dewey number) makes it possible to broaden or narrow a search at will. It is anticipated that the system will be useful to searchers and will also provide a basis for automated indexing. Fifteen computer prototype screens are presented as illustrations. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED 389 310

DEVELOPING AN INFORMATION INFRASTRUCTURE

TO

SUPPORT INFORMATION RETRIEVAL:

TOWARDS A THEORY OF CLUSTERING BASED IN CLASSIFICATION

By

Mary Micco and Rich Popp
Computer Science Department
Indiana University of Pennsylvania
Indiana, PA 15705
412-357-2637

Research Sponsored by:
Council on Library Resources
Department of Education-Library Technology Grant

IR055693

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Helen Mary Micco
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

BEST COPY AVAILABLE

Developing An Information Infrastructure To Support Information Retrieval: Towards A Theory Of Clustering Based In Classification

By

Mary Micco and Rich Popp
Computer Science Department
Indiana University of Pennsylvania
Indiana, PA 15705
412-357-2637

ABSTRACT

The research objective was to explore techniques for building an world-wide information infrastucture by reverse engineering existing databases linking them together in a hierarchical system of subject clusters to create an integrated knowledge base. The controlled vocabulary of the Library of Congress Subject Headings was selected as the thesaurus of choice because of its universal applicability and wide use. This was done to ensure consistency and to group similar items together. Each database is an object in the system covering a given subject area and containing different classes of information packages such as books, articles , reviews, videos etc. Each package(item) within the database is assigned to a subject cluster based on its aboutness, building on information already recorded. An expert system matches the user profile to the information package best suited to their need and then locates the appropriate database. This is supplemented by a machine-generated natural language mapping scheme to lead the user into the clusters of interest. . For this prototype, we developed an object- oriented hypermedia user interface on NeXT workstations with two databases, the first with 100,000 MARC records and the second with 20,000 additional records enhanced with table of contents data. The items are grouped into subject clusters consisting of the class ification number and the first subject heading /keyword assigned. The use of a hierarchical classification number (Dewey) makes it possible to broaden or narrow a search at will. Every other distinct keyword in the MARC record is linked to the subject cluster in an automated natural language mapping scheme which leads the user from the term they entered to the controlled vocabulary of the subject clusters in which the term appeared. The subdivisions of the subject headings are treated as facets and used to limit cluster sizes. The X, Y and Z subfields are indexed separately and used to filter searches by form, place or period. Weighting of terms has been implemented to deal with the problem of very large sets. It is based on extent of contribution to the document. Users can control the specificity by selecting exact match, begins with, appears anywhere. On line subject searching is greatly enhanced by the hypermedia interface. It is anticipated that the system will not only be useful to searchers but can also provide the basis for automated indexing. The indexer will locate items with similar configurations and then can quickly browse to find the best fit from the existing clusters,

I. PROBLEMS IDENTIFIED

Those, who dare to dream, have a vision that some day we will be able to search the word's recorded wisdom over the Internet combining both text and interactive graphics in a multimedia format, browsing through any item or items we select without ever leaving our offices. A number of problems will need to be solved before this vision can become a reality.

1) **Lack of information infrastructure**

There is no infrastructure linking the many thousands of different databases in any organized fashion to facilitate navigation by subject or topic. There are several utilities such as Dialog that offer a collection of these databases with their own proprietary software. They can tell you in which databases a particular word is used, but this is of very little help since no context is provided. We are beginning to see an emerging standard for search software. Z39.50 Information Retrieval Protocol (general) and the more specialized Z39.58 (ref 1) Common command language for online interactive information retrieval. While these standards will do much to assist by providing a common command language they do not address the fundamental issue of of terminology control.

2) **Lack of standards for indexing and vocabulary control**

The lack of these make it difficult to integrate disparate databases in one unified system. In an integrated system we propose that databases should be treated as objects and classified by topics covered, and various other attributes to enable an expert system to match the user with the most appropriate database Within each database a technique for assigning items to subject clusters within each database is proposed.

3).**No distinctionis made between different information packages.**

At present no distinction is made between different information packages. A hit in an abstract of a periodical article is treated the same as a hit in a book or a full text document such as an encyclopedia article or a document in a litigation support system of over 1 million discrete documents of varying sizes. An expert system should match the user profile by need, level of interest, scope of coverage to locate the most appropriate information package before the search for information is initiated.

4). **No context is provided to help the user to sort out the options.**

No context is provided until you are at the actual book level. There are no intermediate levels of abstraction. Psychologists will tell us that the context in which a word is found can be critical in determining its meaning, yet few if any of our current systems provide this except at the document level, e.g. Banks in Finance are very different from Marine Banks which are different again from Sir Joseph Banks the botanist. The context should be derived from the arrangement of the subject clusters in a hierarchical classification of topics.

5) **No weighting of filtering.**

No weighting or filtering is used to assist in obtaining the best of what is available. At the very minimum we should be able to specify the intellectual level of interest. Is it material written for the layman, the professional or the scholar? We should also weight terms by the extent of their coverage in the text. An article that deals entirely with marine banks would be of greater interest than a passing reference in a single paragraph of an article dealing with Japanese fishing rights. We should also be able to request limiters, such as limiting the search to a particular geographic region or a period in history.

6) Little support for navigation to broaden or narrow a search.
Once the search is under way, support for navigation is limited to recording the results of previous searches or the screens the user has visited in a particular database with a particular hypertext software package. Few if any tools are available for helping the user to broaden or narrow a particular search. In database searching where brute force keyword searching has been implemented, users cannot obtain an overview of the topic area or the distribution of the information (fish-eye view).

II. DEVELOPING A WORKING PROTOTYPE

With funding from the Department of Education and the Council on Library Resources, a reasearch project is currently under way at Indiana University of Pennsylvania to explore new algorithms for solving the problems identified utilizing many new technologies including expert systems. The question addressed in this paper is how to succesfully integrate the thousands of existing databases with full text retrieval systems while still being able to locate information on a topic of interest efficiently. The goal is a working prototype that can serve as a model for a much larger experiment in information retrieval across a plurality of databases.

A. SYSTEM OBJECTIVES

The following system objectives have been set

1)Structure the databases hierarchically by topic or coverage.
Each database is treated as an object in the system with its special attributes. These are classified hierarchically according to their subject matter, while also recording their depth of coverage , type of material, intended audience , organization and availability of a thesaurus..

2) **Develop an Expert System to match the user with the information package.**
An expert system component has been developed to match the user profile with the information package best suited for the information need identified. We decided to use the 25 basic document types defined in the DOD standard(Ref2) . Clearly it would be very helpful to be able to restrict a search to particular document types deemed appropriate or to exclude those that were not of interest. With this in mind, we were able to specify document types as one of the attributes for the databases in our system. In this way the expert system could direct the user to the databases most likely to include the information packages best suited to their request.

3.) **Create Subject Clusters:**
Within each database materials are gouped into subject clusters based on their "aboutness " as it relates hierarchically to the database topic as a whole. The purpose is to collect similar items together into clusters which can then be grouped in super clusters. In this way we can offer different levels of abstraction depending on the need. The user can examine the clusters, broadening or narrowing their search at will to locate information of interest. They can also quickly determine how the material is distributed in a general topic area in much the same way as road maps. Currently the traveller who wishes to drive from San Francisco to New York will only be interested in the major interstate throughways. When reaching New

York State they may require more detail such as the major 4 lane highways. But when they reach New York they will want a detailed street map to help them find their destination. We should offer similar guides to our databases. Conversely one might enter a search for Cedar St and be told that 300 cities have a street named Cedar St. At this point the user would want to ask for additional contextual information to help to choose. This system of developing subject clusters automatically will be applied uniformly across all the databases. The subject clusters consist of a numerical classification number representing the concept as it relates hierarchically to other materials in the database as well as an additional controlled heading chosen from a thesaurus. Since each classification number is represented by an English caption, we have chosen to link these in as an additional source of keywords for searching. To prevent a combinatorial explosion, each item is assigned to one and only one subject cluster.

4) **Search for subject clusters instead of keywords.**
Apart from the useful context that the subject cluster provides, there is the added advantage of grouping similar items together so that in a large set the user can quickly see the major groupings and decide which ones are worth further exploration.

5) **Impose a weighting scheme**: It is very clear that some sort of weighting scheme needs to be developed in very large retrieval systems because simply retrieving all the clusters in which a given word appears frequently generates very large sets. The average number of items retrieved in succesful searches in Melvyl, a database ofover 6 million items , is over 171 documents, far more than any user is prepared to deal with. We decided in our system to experiment with 5 different levels of weighting for keywords found in full text documents based on the ISO standard formats for documents as follows.
.Z39.2 MARC Machine Readable Cataloging format. This is a standard document header. The format is based on tagging key elements in the record. Found principally in the library world, it has also been used for various litigation support systems where a header with keywords is generated and then the documents themselves are microfiched in full text. The principal advantage of developing such a header for each document is to aid in database management. But another benefit is that key elements of the document are organized in such a way as to facilitate building separate indexes as well as weighting. The format is very flexible with a great variety of tags to choose from, most of which can be used more than once as needed. If a controlled vocabulary is being used to identify the main topics treated in the document the 600 tags are used. These can also be used for a document abstract or summary of the main headings or table of contents. A list of keywords can be included if desired. In our weighting scheme , we chose to consider the first 600 tag assigned as the 'aboutness' of the document. with a Weight 1. Other controlled terms (600's) were considered as Weight 2, whereas natural language terms drawn from the title, table of contents and/or abstract were considered Weight 3. From there we moved to the full text of the document and decided to take advantage of the SGML markings for subtitles, chapter, section and paragraph headings.(Z39.59 SGML Standard Generalized Markup Language) All of these were considered as Weight 4. All remaining keywords were given a weight of 5. By taking advantage of the existing standards and current practices we were able to implement a weighting scheme with very little additional effort.

6) Use filters to narrow large sets:

It was also important to take advantage of the use of filters. Traditional filters commonly supported involve personal names, language of the document and date of publication or distribution. Other useful options tested in the IUP project involved period, place, document type or form, and general facets of a topic eg. legal, economic, social. Typically, certain facets of a document warrant special indexing because of their usefulness in retrieval. If these are assigned special tags then it is relatively simple to index these separately thereby enabling the searcher to limit a large set to documents dealing with a given period, or place for example. Another option, more easily added, is to add on these options as facets of a given topic. Each facet is treated as a separate subfield and the MARC specification referred to earlier provides special subfield delimiters for specific facets. Typically, the y or d subfield delimiter is used for dates, the z subfield is used to indicate place.

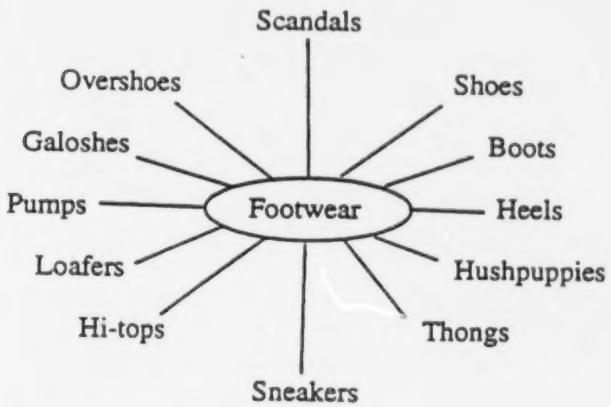
B. FORMING THE SUBJECT CLUSTERS USING CLASSIFICATION

1. Theoretical Background

For many years, since the Cranfield Studies, researchers have maintained that brute force keyword searching in databases is the most effective information retrieval technique in spite of many efforts to improve the precision of searches by a variety of roles, links, and facettting schemes. The problem now is that many systems today are becoming so large that it is not uncommon to retrieve hundreds of items when searching for all occurences of a keyword in the database. For searches in Melvyl the average set size is over 170 items, while common words such as United States, international and war yield search sets of tens of thousands. There is a real need to develop techniques for dealing with very large sets. These searches lack precision since there is no way to determine the context in which the word is used except to check the item entry itself.

2 Brute Force Keyword Searching

Another serious problem but a much harder one to document, is the question of what is being missed. Consider for example the many words we use to describe footwear.



In a large periodical database each of these words will probably yield a few hits but the user will have no hint that other related material is available.

3. Use of controlled Vocabularies/Thesauri.

The solution proposed has been to use a controlled vocabulary derived from a thesaurus to cluster similar items together. For instance in this case one might choose the controlled heading 'Footwear'. By itself it will be of little value as very few authors and even fewer users would be likely to choose this term. In Melvyl a large library database of over 6 million items they are reporting over 30 % failed searches (0 hits) using the controlled vocabulary of the Library of Congress subject headings. Some other systems have experimented with offering both keyword searching and controlled vocabularies. The following search for material on the foreign relations between the Soviet Union and the United States was submitted to a 500,000 document database managed by software from Carlyle Systems which offers both a controlled vocabulary and keyword searching with the following results.

4. Combining controlled vocabularies with keyword searching.

Foreign relations 4265 (controlled heading)
.....and United States 2394 (controlled heading)
and Soviet Union and United States 75(controlled heading)
Uncontrolled keyword searches:
.....and USA 0 hits
and US 102 hits
and USSR 24 hits
and Russia 432 hits
and Soviet Union 189
and USSR 0
and US and Soviet Union 0

There are currently a number of systems which offer both forms of searching but once again the results are not very satisfactory as can be seen from the following search. Subjects were asked to develop a search strategy for finding material on the foreign relations of the US with the Soviets.

Using the controlled vocabulary there were 2394 items dealing with the foreign relations of the United States, 75 of which also dealt with the Soviet Union. The results show clearly the variety of strategies that can be developed. Most revealing is the fact that of the 73 unskilled users who attempted this search only 8 stumbled onto the controlled heading. The rest found some material but not necessarily the best and they left not realizing that they had missed a large percentage of what was in the data base. Even more disturbing was the fact that the users who had 0 hit searches left the system convinced that there was nothing available. They made few if any efforts to alter their initial strategies.

5. Need for a Natural language mapping scheme.

Our search failure analysis pointed to a need for a dense semantic network linking the users' natural language terms to the controlled subject headings.. In this way we would have the best of both worlds. The user would be guided from the term they input to the controlled headings of the items in which the word occurred. Each controlled heading represents a

subject cluster of similar materials. For example the person searching for 'sneakers' is reminded that the controlled term is 'footwear' and is led to this subject cluster. The subject cluster for footwear contains all the various subtopics in items about footwear. At the same time the indexing is handled in such a way that the user can still find the items with specific references to sneakers if this is really what they want. This dense semantic network was accomplished without any human intervention. Every term and bound phrase in the document is linked to the subject cluster for that document, its 'aboutness' whether it be a full text document of multiple pages or a two line author/title entry. These terms are indexed in a central dictionary. The user will be guided from the term they input to all the subject clusters in which the word appeared. Altogether we developed 1,346,267 distinct keywords or over 30 per cluster working only with the document headers not the full text.

5 Assigning items to clusters: Automating the process.

In our prototype, with 100,000 records we experimented with techniques to generate the subject clusters and the natural language links automatically. The key issues were:
a). The ratio of subject clusters to the size of the database. The experimental system was designed to scale up to databases of several million documents. This being the case we set a hard and fast rule that each document in the system could belong to one and only one subject cluster, chosen to represent as closely as possible the 'aboutness' of the document. Every other term in the document would be linked to the subject cluster and would serve as a possible lead in. Cluster size is monitored continually as new documents are added and if the number of documents in a cluster is over 50 it is flagged for human intervention. It can then be subdivided into narrower topics.
b).Assigning books to clusters automatically.
Defining the 'Aboutness' of the document.
We experimented with a variety of techniques for forming the subject clusters automatically by building on information already contained in the record. Since we wanted to demonstrate the feasibility of 'reverse engineering' many existing databases we were also careful to consider the many standards already developed. The 100,000 records we used for our prototype were provided by the Australian Defence Force Academy Library, and utilize the MARC (MAchine Readable Cataloging record format defined in Z39.2)

001 172975
008 880023s1987 nju . eng
043 00 aaw—
043 00 aa-cc—
043 00 aa-kr—
082 00 a355.4 b2 b0904
100 10 aFlint hRoy K. kRoy Kenneth c1928-
245 14 aThe Arab-Israeli wars, the Chinese Civil War, and the Korean War dRoy K. Flint, Peter W. Kozumplik, Thomas J. Warnka
260 00 aWayne, N.J bAvery Pub. Group cc1987
650 00 aKorean War, 1950-1953
650 00 aMilitary art and science xHistory y20th century
651 00 aIsrael xHistory, Military
651 00 aArab countries xHistory, Military
651 00 aChina xHistory yCivil War, 1945-1949

Example of Marc Record Format

This very flexible format has been widely accepted not only in libraries but also in a variety of records management applications particularly litigation support systems. It depends on a series of optional tags which can be repeated as many times as needed to fully describe the document. and has been used on occasion in full text systems to provide a header for the text itself.

We studied a number of options to create the subject clusters from the Marc records.

1) Using titles assigned by the authors.

This proved to be unsatisfactory because it resulted in generating 95130 different clusters for 100,000 documents. Both authors and publishers strive to make their titles unique.

2) Using the controlled terms assigned:

Since all of the documents in our collection had been assigned one or more controlled headings from a thesaurus (Library of Congress Subject Headings), we established a heuristic that the first heading assigned represents the 'aboutness' of the document. and is the primary heading. This reduced the number of clusters to 43,663 per 100,000 but did not completely solve our problem. We found that over 90% of the clusters had 5 or less documents while the remaining 10 % of the clusters became vey large. Another disadvantage was the limited vocabulary of the thesaurus, some 30,000 keywords, many of which were very general in nature and rather stilted. It was very difficult to use the thesaurus to broaden or narrow a search since the syndetic structure was weak and tended to lead the user away from the topic to other broad related areas (Sinkankas). Another problem was that where more than one subject heading had been assigned, we had to distinguish those headings that were being used as subject clusters and those that were not. This difficulty led us to the development of a weighting scheme (see below).

3) Using the classification number.

As an alternative we experimented with using the classification number to cluster the documents on line since these are currently used to group similar items together on the shelves. Each item is assigned one and only one class number denoting where it fits in the universe of knowledge. This method yielded 28,951 clusters per 100,000. If we chose we could abbreviate the clusters to the first 3 digits resulting in 915 clusters. For periodical articles the classification number used was drawn from the periodical itself.

There are two major universal clssification schemes in use for organizing information. The first is organized hierarchically from the top down..the Dewey Decimal Classification System and its cousin, the Universal Decimal Classification Scheme. There are captions for each number and these can provide additional keywords for searching. The second system the Library of Congress Classication System is a bottom up system that takes each item as it comes and positions it on the shelf in relation to the items that are already there. Unfortunately the notation used is not hierarchical nor does it reflect the arrangement of the subjects in relation to each other. There is no index. The captions used for the numbers are not yet available in machine-readable form and so this system was abandoned as impractical. The great advantage of using a classified approach was that the subject clusters could then be linked in a hierarchically organized tree of knowledge enabling the user to broaden or narrow his search at will. A search for gold lead us to the cluster in the hierarchy shown below among others and we can quickly see how the database is structured and where gold fits under money in the general topic of economics. If we were interested only in mining gold, then we could immediately eliminate this cluster.

300 Social Sciences
330 Economics
.1 Banks and banking
.2 Specialized banking institutions
.3 Credit and loan institutions
-.32 Savings and loan institutions
.4 Money
.4042 Gold and silver
5 Other mediums of exchange
.6 Investment finance
.7 Credit
.8 Interest and discount
.9 Counterfeiting,forgery, alteration.
333 Land economics
334 Cooperatives
335 Socialism and related systems
Dewey example.

4) A hybrid system.

We expected to find a one to one relationship between the subject heading assigned and the classification number/caption. To our suprise this proved not to be the case. Each subject heading had the potential of one or more classification numbers , while conversely each classification number had more than one subject heading, a classic many to many relationship.

General History/Central Europe/Germany/1866+

Germany—Politics And Government—1933-1945	943.08	2
Nationalsozialistische Deutsche Arbeiter-Partei—Schutzstaffel	943.08	2
Germany—History—1933-1945	943.08	4
Hitler, Adolf—1889-1945	943.08	17
National Socialism—Terminology	943.08	1
United States—National Archives And Records Service	943.08	1
Hess, Rudolf—1894	943.08	1
Bormann, Martin—1900	943.08	1
Neurath, Konstantin Hermann Karl—Freiherr Von—1873-1956	943.08	1
Moltke, Helmuth James Wilhelm Ludwig Eugen Heinrich—Graf Von	943.08	1
Hitler, Adolf—1889-1945—Assassination	943.08	2
Hitler, Adolf—1889-1945	943.08	17
Germany, West—History	943.08	1
Germany Federal Republic, 1949	943.08	1
German Reunification Question 1949	943.08	1
Germany—History—Allied Occupation, 1945—Sources	943.08	1
Germany Territory Under Allied Occupation, 1945-1955 French Zone	943.08	1
Adenauer, Konrad—1876-1967	943.08	1

This was readily solved in the customary manner by identifying our clusters by the unique combination of subject heading and class number. As expected this increased the number of clusters, resulting in 83,551 per 100,000. We have not yet had the opportunity to test our clustering method on a database of millions of records but we are predicting that the sharp growth in the number of new clusters will taper off and cluster sizes will increase noticeably as the database grows.

Both the primary subject heading and the class number for the relevant subject cluster are displayed when an item is retrieved as well as the English caption for the class number to create a context and to provide information about where the item fits in the general order of knowledge. The value of the context in increasing the precision of searches and helping users to rapidly eliminate noise cannot be overemphasized and is supported by a considerable body of research studies in the psychology of learning. People can recognize far more than they can recall. The display method chosen builds on their findings. Instead of listing all retrieved items in reverse chronological order of acquisition which is the current practice, we display only one line of information about each subject cluster found including the number of relevant items These subject clusters can be sorted hierarchically by their classification number, or alphabetically by subject heading at the users' discretion.

The following is an example of the subject clusters found by a search for 'banks' sorted by class number.

Financial economics	Underdeveloped Areas--Banks And Banking	332	1
Cooperatives	Banks And Banking, Cooperative--United S	334	1
Production	Blood Banks--Economic Aspects--United St	338	1
Admin of US federal/state gov.	Banks And Banking, Central--Us	353	1
Pharmacology and therapeutics	Blood Banks	615	1
Hunting/fishing/conservation technologies	Fisheries--Grand Banks Of Newfoundland	639	1
Accounting	Banks And Banking--Accounting	657	1
General management	Mortgage Banks--United States	658	1
Advertising and public relations	Public Relations--Banks And Banking	659	1
English/Old English (Anglo-Saxon) literatu	Banks And Banking--Great Britain	820	1
English/Old English (Anglo-Saxon) literatu	Land Banks--Great Britain	820	1

As can be seen there is only one line for each cluster but a great deal of contextual information is being provided to assist the user in increasing the precision of their search. It is much easier to judge the relevance of a document if it is shown in its context. Searching for a keyword produces a list of all the subject clusters in which the term appeared. Users can then select the cluster or clusters of interest.

C. USING A SYSTEM OF WEIGHTS AND FILTERS TO NARROW LARGE SETS

Since a picture is worth a thousand words, we have included a search for material on the United States, a particularly large set, to demonstrate the value of weighting, filtering and truncation in managing large retrieved sets. A brute force key word search for **united states:** accepting only entries that begin with the phrase United States, yields the following:

Books: 13,905 Subject Clusters 3,762

1. Using facets to narrow a search.

a) Search for United States using Weighting and facetting.
When we requested instead a Display of only Weight 1 books in which United States appears as the primary subject heading, we reduced the number of books to 15% of the original total while the subject headings represent 30% of those found in the brute force approach.

Books: 2096 Subject Headings: 1151.

Each different arrangement of facets or subdivisions in the heading is counted as a separate heading. Note: United States--Congress--House (2 facets) is considered to be different from United States--Congress--House --Committees(3 facets). The value of presenting contextual information for each cluster in one line cannot be overemphasized.
The user can quickly scan the topic area to learn what different facets are being treated and how the literature is distributed. For instance it may be useful to know that there are Directories of the US Congress as well as several Biographies and a Bibliography.
Questions have been raised for years about the use of facets. Ranganathan proposed a highly structured scheme but it has never been widely used because the high overhead and complexity of the scheme were not cost effective. The Library of Congress has been assigning subject headings as needed from a very large group of acceptable headings. This system seems to work well as it is fairly simple to implement and very effective in breaking up large sets into manageable subgroupings.

Number Of Books Found:2096; Number of Subject Headings Returned:1151

SUBJECT HEADING			
United States--Commerce--History	658	General management	1
United States--Commerce--History--Source	382	International commerce (Foreign trade)	1
United States--Commerce--Russia	382	International commerce (Foreign trade)	1
United States--Commerce--Spanish Americ	382	International commerce (Foreign trade)	1
United States--Commerce--West Indies, Br	382	International commerce (Foreign trade)	1
United States--Commercial Policy	338	Production	1
United States--Commercial Policy	382	International commerce (Foreign trade)	1
United States--Comptroller Of The Curren	353	Admin of US federal/state gov.	1
United States--Congress	328	The legislative process	9
United States--Congress--Bibliography	016	Bibliographies & cats of works on specific s	1
United States--Congress--Biography	328	The legislative process	3
United States--Congress--Conference Comm	328	The legislative process	1
United States--Congress--Directories	328	The legislative process	1
United States--Congress--Elections	329	[Unassigned]	1
United States--Congress--House	328	The legislative process	3
United States--Congress--House--Committe	328	The legislative process	1

b). Same Search Limited to Class Number and One Facet only

One of the difficulties with using subject headings to cluster the books is that 80% of the facetted clusters are less than 5 books even in databases of 7 million records (Melvyl). As we worked on how to create larger book clusters it occurred to us that the software could collapse the facets into larger clusters. Here we have collapsed United States to one facet for each heading. Note cluster sizes have increased. United States--Politics And Government--973 yields a cluster of 48. . The number of books remains constant at 2096, but by clustering the facets we have reduced the number of Subject headings from 1151 to 554 or only 15% of the original total.

Number Of Books Found:2097; Number of Subject Headings Returned:554

United States--Politics And Government	324	The political process	1
United States--Politics And Government	327	International relations	1
United States--Politics And Government	328	The legislative process	7
United States--Politics And Government	329	[Unassigned]	5
United States--Politics And Government	332	Financial economics	1
United States--Politics And Government	342	Constitutional and administrative law	2
United States--Politics And Government	353	Admin of US federal/state gov.	8
United States--Politics And Government	355	Military science	2
United States--Politics And Government	909	World history	1
United States--Politics And Government	917	Geography/travel in North America	5
United States--Politics And Government	940	General history of Europe/Western Europe	1
United States--Politics And Government	973	United States	48
United States--Politics And Government	975	Southeastern U. S. (South Atlantic states)	1
United States--Politics And Government	980	General history of South America	1
United States--Popular Culture	001	Knowledge	1
United States--Popular Culture	301	Sociology and anthropology	1

c). Same Search Limited to Class No (no facets)

Here all facets are subsumed in one cluster with the effect that the heading is United States and the user is depending on the classification number to break down the set in a useful way. We have now reduced the number of subject headings returned to 163 or 4% of the original total of 3762. Note the material is arranged in the logical order of the Dewey classification system. The user can broaden or narrow the search at will, massaging the data to explore what is available or to find alternatives to their current search strategy.

Number Of Books Found:2097; Number of Subject Headings Returned:163

CLASSIFICATION	SUBJECT-HEADING	DEWEY (973)	TITLES
Persons in social sciences	United States	923	1
Persons in the arts and recreation	United States	927	1
Genealogy, names, insignia	United States	929	3
General history of Europe/Western Europe	United States	940	16
England and Wales	United States	942	1
Central Europe Germany	United States	943	1
Iberian Peninsula/Spain	United States	946	1
China and adjacent areas	United States	951	1
Southeast Asia	United States	959	2
General history of North America	United States	970	3
United States	United States	973	420
Northeastern U. S./(New England/Mid- Atla	United States	974	2

Example of a search with all facets collapsed

2. Using facets to broaden a search.

a) Expanding a Particular Cluster

User selects a cluster United States--973 and expands it to see the first facet of each heading. There were 420 books in this cluster. In the breakdown another particularly large cluster shows up 973 -- United States--History.

Number Of Books Found:420; Number of Subject Headings Returned:40

CLASSIFICATION	SUBJECT-HEADING	DEWEY (973)	TITLES
United States	United States--Emigration And Immigratio	973	1
United States	United States--Foreign Relations	973	22
United States	United States--Hist X Civil War	973	1
United States	United States--Historic Houses, Etc	973	1
United States	United States--Historiography	973	7
United States	United States--History	973	241
United States	United States--History, Military	973	7
United States	United States--Intellectual Life	973	7

b) Double Expanding a Particular Cluster

User expands again on a cluster United States--973--History and obtains a breakdown of all the facets in this set as shown. In this manipulation of the facets no distinction is made between different kinds of facets. They are taken in the order in which they were assigned. Another approach , the use of specific tags to narrow large sets is shown below.

BEST COPY AVAILABLE

Number Of Books Found:240; Number of Subject Headings Returned:104

			TITLES
United States	United States--History--1945	973	2
United States	United States--History--1945--Sources	973	1
United States	United States--History--1961-1963--Miscell	973	1
United States	United States--History--1961-1969	973	1
United States	United States--History--20th Cent	973	1
United States	United States--History--20th Century	973	2
United States	United States--History--Addresses, Essays, I	973	1
United States	United States--History--Caricatures And Ca	973	1
United States	United States--History--Chronology	973	4
United States	United States--History--Civil War	973	2
United States	United States--History--Civil War, 1861-18(	973	10
United States	United States--History--Civil War, 1861-18(	973	1

A Particular Cluster United States-973 Expanded again

3. Using Filters By Form, Period and Place To Restrict a large set

Current search software generally does not support filtering sets except by the major tags such as author, publication date and language. These are very crude breakdowns. We have found that the controlled headings used by Library of Congress or other formally structured thesauri are not well known to the general public, with the result that they do not know how to effectively use subfields to break up a large set. To alleviate this problem we took advantage of the subfield codes provided in the MARC record format and extracted various specialized subfields as follows

x.....general subdivision
y.....period
d.....date (ususally for biographies)
z .....place

These were sorted into separate indexes accessible through special pop-up windows. This made it possible for us to offer the choice of 4 new limiters: period, place, form and general topical subdivisions.

a) Form Divisions.
We simply sorted through the x subfields and extracted all headings that seem to represent a form of the material e.g. history, bibliography,biography, fiction, periodicals, catalogs , etc. The user's search results are 'and'ed' with this index and all the matches displayed. The user can quickly determine which form headings have been used with their topic. In this case they selected United States --History and form addresses, essays,lectures with the following results. Several of the users suggested that speeches was the term they were expecting to find.

BEST COPY AVAILABLE

Number of Subject Headings Returned:51 form: Addresses, Essays, Lectures

SUBJECT HEADING	DEWEY (abbr.)	CLASSIFICATION	TITLES
United States--History--Addresses, Essays, I	973	United States	1
United States--History--Civil War, 1861-18(	973	United States	10
United States--History--Civil War, 1861-18(	973	United States	2
United States--History--Civil War, 1861-18(	973	United States	5
United States--History--Revolution 1775-17	973	United States	1

b) General or Standard Subdivisions

After eliminating the frequently used form subdivisions (used more than 100/100,000 times) we found 45 other topical subdivisions that were generally applicable)
The user can browse through these in a pop up window and select any of interest such as politics and government, social conditions, description and travel, foreign relations, etc.
In the search shown the user selected Commerce as the subject of interest.

Number of Subject Headings Returned:19 general: commerce

SUBJECT HEADING	DEWEY (abbr.)	CLASSIFICATION	TITLES
United States--Commerce	338	Production	1
United States--Commerce	650	Management and auxiliary services	1
United States--Commerce	658	General management	1
United States--Commerce--China--History	382	International commerce (Foreign trade)	1
United States--Commerce--Directories	380	Commerce, communications, transportatior	1
United States--Commerce--Europe, Eastern	382	International commerce (Foreign trade)	1
United States--Commerce--Germany, West	382	International commerce (Foreign trade)	1
United States--Commerce--History	658	General management	1
United States--Commerce--History--Source	382	International commerce (Foreign trade)	1
United States--Commerce--Russia	382	International commerce (Foreign trade)	1
United States--Commerce--Spanish Americ	382	International commerce (Foreign trade)	1
United States--Commerce--West Indies, Br	382	International commerce (Foreign trade)	1
United States--Economic Conditions	330	Economics	12

c) Period

We extracted the d and y subfields (used for dates) and created a separate index with two additional fields, begin year and end year. The user is prompted for the begin and end year of the period they want and the proper search strategy is developed and submitted for them utilizing the period index. It was clear as we did this that the usage of the d and y subfields was not always consistent and that we missed numerous headings, but this is a data entry problem and could be corrected at that level. We also provided for the user who simply wanted to know what we had in a given period such as the Middle Ages. The system can either return all materials for a given time period or narrow a given set down to materials within a time period.

BEST COPY AVAILABLE

Number of Subject Headings Returned:87, Period: 1935 - 1950

SUBJECT HEADING	DEWEY (abbr)	CLASSIFICATION	PERIOD
United States--Foreign Relations--1977-198	327	International relations	1945
United States--Foreign Relations--1981	327	International relations	1945
United States--Foreign Relations--China	327	International relations	1949-
United States--Foreign Relations--China	940	General history of Europe/Western Europe	1937-1945
United States--Foreign Relations--Greece	327	International relations	1944-1949
United States--Foreign Relations--History	353	Admin of US federal/state gov.	1945
United States--Foreign Relations--Latin An	327	International relations	1948
United States--Foreign Relations--Russia	327	International relations	1945
United States--Foreign Relations--Soviet U	327	International relations	1945
United States--Foreign Relations--Spain	327	International relations	civil war,
United States--Foreign Relations--Spain	327	International relations	1939
United States--Foreign Relations--Spain	946	Iberian Peninsula/Spain	civil war,
United States--History--1933-1945	917	Geography/travel in North America	1945
United States--History--1945	973	United States	1945

Example of United States materials in a given period: 1935-1950

d) Place

This was by far the most frequently used form of subdivision. In Melvyl over 70% of all subdivisions used involved place names. In the Marc record format there are two ways in which geographic data can be entered. A formal tag, 042 has been dedicated for geographic area codes. A formal list of these is in place and the English names of the codes has been provided. We also found that most subject headings are subdivided by place as a facet when appropriate. For the user they can enter the place name as a keyword and search for all occurences or they can search for a topic such as foreign relations and limit it by country or area such as United States. .

For instance they might be interested in any interchange that involved the United States and China. Unfortunately the system can only retrieve the material that is indexed. We have found many cases where the geographic area codes(gac) were not used even though the subject heading indicates that the material involves the country. We have been experimenting with building an index that combines both the gac codes and the z subfields but in reality this should be worked out on input. Another problem is that the most specific geographic unit is used in the subject heading e.g Potts County but there is no connection made to Pennsylvania.. Unless a gac code is also assigned, this leaves us with no mechanism for exploding a broader heading to capture the smaller units within that area. The user who asks for material on Pennsylvania will not retrieve Potts county. These problems can readily be resolved by making more effective use of a classified system of geographic names. References to China appear in other interesting clusters such as United States--Foreign Relations--Japan

Number of Subject Headings Returned:16 Area: China

SUBJECT HEADING	DEWEY (abbr.)	CLASSIFICATION	TITLES
United States--Commerce--China--History	382	International commerce (Foreign trade)	1
United States--Commerce--China--History	382	International commerce (Foreign trade)	1
United States--Foreign Relations--1953-196	973	United States	7
United States--Foreign Relations--China	081	American	1
United States--Foreign Relations--China	327	International relations	14
United States--Foreign Relations--China	327	International relations	14
United States--Foreign Relations--China	327	International relations	14
United States--Foreign Relations--China	327	International relations	14
United States--Foreign Relations--China	940	General history of Europe/Western Europe	1
United States--Foreign Relations--China	940	General history of Europe/Western Europe	1
United States--Foreign Relations--China--S	327	International relations	1
United States--Foreign Relations--China--S	327	International relations	1
United States--Foreign Relations--Japan	973	United States	1
United States--Foreign Relations--Taiwan	327	International relations	1
United States--Foreign Relations--Taiwan	327	International relations	1
United States--Foreign Relations--Taiwan	327	International relations	1

Example of search for United States limited to China

4. Use Of Truncation To Limit Sets.

Another very effective method of managing large sets is to implement truncation for the user. We have found three different levels of truncation in subject cluster names to be useful. The first 'Exact Match' requires that there be an exact match with the phrase presented by the user. In the case of United States, such a strategy would only retrieve items that dealt with the United States in general terms. The second option, 'Begins with', has been designated the default strategy and will match any subject cluster name that begins with the phrase input by the user as seen in the searches for material on the United States and all its facets. The third option ' Appears anywhere' will find any subject cluster where the United States is mentioned even if only as a geographic subdivision. e.g. Swordplay--United States
The following search on international demonstrates the value of the truncation options most effectively. A brute force search for internation yields 3975 books with 1150 subject headings. The searches shown below were done with only Weight 1 materials. The word international had to appear in the primary subject heading assigned to the item.

a) **Exact Match** on International

Number Of Books Found:1; Number of Subject Headings Returned:1

SUBJECT HEADING	DEWEY (abbr.)	CLASSIFICATION	TITLES
International	335	Socialism and related systems	1

BEST COPY AVAILABLE

b) **Begins With:** The same search done with the Begins With option retrieves 253 items with 120 headings as shown.

Number Of Books Found:253; Number of Subject Headings Returned:120

SUBJECT HEADING	DEWEY (abbr.)	CLASSIFICATION	TITLES
International Agencies--Addresses, Essays,	341	International law	1
International Agencies--Bibliography	016	Bibliographies & cats of works on specific s	1
International Agencies--Directories	016	Bibliographies & cats of works on specific s	1
International Bank For Reconstruction An	332	Financial economics	1
International Brotherhood Of Teamsters, C	331	Labor economics	1
International Brotherhood Of Teamsters, C	331	Labor economics	2
International Brotherhood Of Teamsters, C	331	Labor economics	1
International Business Enterprises	331	Labor economics	1
International Business Enterprises	338	Production	16

c) Appears Anywhere:

Can be very useful if the user is unsure of the heading but knows the word international is in there some place

Number Of Books Found:372; Number of Subject Headings Returned:182

SUBJECT HEADING	DEWEY (abbr.)	CLASSIFICATION	TITLES
American Federation Of Labor And Congre	331	Labor economics	1
Amnesty International	323	Civil and political rights	1
Arbitration, International	172	Political ethics	1
Arbitration, International--Congresses	327	International relations	1
Arbitration, International--Digests	341	International law	1
Astronautics--International Cooperation	341	International law	1
Astronautics--International Cooperation--C	629	Other branches of engineering	1
Bank For International Settlements	973	United States	1
Banks And Banking, International	332	Financial economics	4
Berlin--International Status	341	International law	3

IV. CONCLUSION

It is very clear that the new technologies can be harnessed effectively to make major improvements to the management of our online information resurces without the need for a massive conversion effort. We have taken the elements that already exist in the records and manipulated them in new and interesting ways to provide better subject access -- a form of reverse engineering if you will. In the process we have demonstrated once again the value of classification to manage an infrastructure and to assist navigation by enabling the user to broaden or narrow searches at will. We explored the relationship between the class number

BEST COPY AVAILABLE

or parking place assigned to the item and the controlled subject headings or keywords assigned to capture the various topics. We took advantage of this to create subject clusters of similar items instead of the statistical clustering favored by so many.

An additional benefit of applying our techniques to the massive databases that are already out there, is that we have instantly created a system which can help indexers to see the distribution of the literature and quickly determine the best fit for the item being indexed. An organized and coherent information infrastructure based on classification with standardization in the techniques for developing subject clusters is absolutely essential if we are to provide the level of service of our users have come to expect.

The project is still under way and we expect to begin serious user studies in the near future to determine how users react to the many changes. However our research will focus on features of the interfaces that need to be modified or improved to support searching. It is quite clear that searching and displaying subject clusters, the use of weighting, filters and facetting all are badly needed improvements to brute force keyword searching when very large sets are retrieved.

The research question is not whether change is needed but how to make the changes most effectively. Which classification system is ultimately selected is difficult to predict but our research was intended to explore the possible use of current systems and to develop some specifications.

BIBLIOGRAPHY

Alberico, Ralph and Mary Micco. *Expert systems for reference and information retrieval Westport,* CT: Meckler Publishing Co, 1990.

Arms, Caroline, Editor. *Campus Strategies for Libraries and Electronic Information. Bedford,* Mass: Digital Press, 1990 (EDUCOM Strategies Series on Information Technology)

Bates, Marcia. (1986) "Subject acess in online catalogs: a design model". *Journal of the American Society for Information Science* (37:6) (January 1986) 357-375.

Crawford, Walt. *Technical Standards: An Introduction for Librarians.* Professional Librarians Series. White Plains, N.Y.: Knowledge Industry Publications, 1986.

Lynch, Clifford. In Hudson, Judith and Walker, Geraldine. "The year's work in technical services research, 1986". *Library Resources and Technical Services* (December, 1987) 275-286.

Markey, Karen and Visine-Goetz, Diane. "Increasing the accessibility of Library of Congress subject headings in online bilbiographic systems". *Annual Review of OCLC Research,* 1988 (1987-88) 32-34.

Settel, Barbara and Cochrane, Pauline."Augmenting subject descriptions for books in online catalogs", *Database* (December, 1982)